Coursa Capstone Project:

The Battle of Neighborhoods – the case of Berlin

1. Introduction

I choose the city of Berlin. Here I want to discuss the issue of rental housing prices.
Berlin is considered to be one of the most popular and hippest cities both nationally and internationally. It offers many possibilities both for working and living. Last but not least, the diverse cultural offerings and the exciting nightlife attract many to the city. The cost of living is also relatively low in comparison to other metropolises around the world and makes Berlin even more attractive, so that young people are increasingly drawn to the city. Nevertheless, in recent years it has become apparent that not only has the cost of living risen in the city, but above all affordable housing is becoming increasingly scarce. As in many other metropolitan areas, rents for apartments in Berlin have risen in recent years and there is no end in sight, so that young people who come to Berlin as career starters are looking in vain for accommodation in the hip districts.

If one considers the average rent of the individual districts of the city, one also finds that the residential location is not necessarily related to the co-prices and that there must therefore be other factors that justify a district as attractive and thus the rent. Supposedly hip districts are particularly expensive while pure residential districts are pushed into the background.
This project therefore examines rental prices in Berlin and its individual districts and investigates the question: What are hip districts with expensive rents characterised by what venues? What are the top venues in an expensive district as that consider hip and ‚fancy‘.

2. Dataset

To solve the problem, we will need the following data

- List of neighbourhoods in Berlin
- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data: Sources of data is Wikipedia which contains a list of neighbourhoods in Berlin. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods
- Venue data, particularly data related to the districts: We will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+

million places and is used by over 125,000 developersto explore the neighbourhood regarding their top 10 venues. Using the Foursquare data in order to characterize the popular and expensive districts to determine.

- Resident location: good, medium, simple: These variables indicate to what extent a district within the city is attractive as a place to live and one could assume that districts with good residential locations tend to have expensive rents.
- Population density: The more inhabitants, the less space and the more expensive living space.
- Living space: The more living space, the more expensive the rent (whereby it is neglected here that the living space is related to how large a district is and over how much space is available)
- Number of residential buildings: The more residential buildings, the less the rent should be.
- Distanace from city centre: using geopy coodrinates and calculating eacht districts distance to city center

3. Method /Approaches

These assumptions will be evaluated through correlation and regression analysis so that appropriate variables can be selected to cluster the districts.

This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

Firstly, I need to get a list of Berlin City's neighbourhoods in the city of Kuala Lumpur, which is available through a wikipedia page .I will do web scraping using Python requests and use pandas dataframes to extract the list of neighbourhoods data. However, this is just a list of names. Moreover will I need to get the geographical

coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, I will use Geocoder package that will allows me to convert the districts names into geographical coordinates in the form of latitude and longitude. After gathering the data, I will bringt the data into a pandas DataFrame and then visualize the the results in a map using Folium package. Parte of the datat will be the amount of venues in a given district provided by Foursquare data. I will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters.

We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. The result will be the a number of venues in each district which I will include in the Dataframe to do a regression analysis.

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. The results will allow us to identify neighbourhoods along their distribution of venues. Based on the structure of venues in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most vivid and might explain why rental prices in those districts are high.

4. Exploratory Analysis

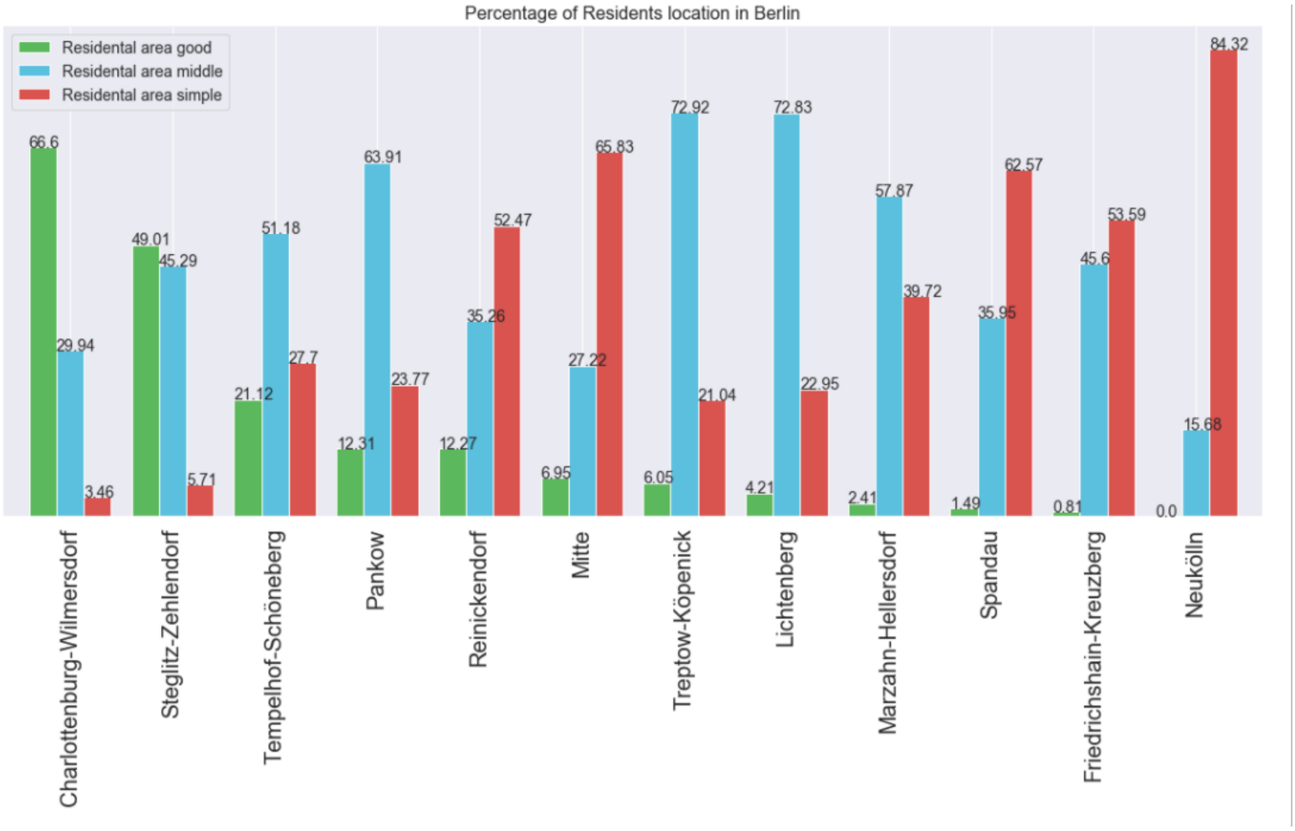
First step of the exploratory analysis was the creation of the dataset itself via different sources. This Wikipedia page (https://de.wikipedia.org/wiki/Liste_der_Bezirke_und_Ortsteile_Berlins) contains a list of neighbourhoods in Berlin. I will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and pandas dataframes. Afterwards I will get the geographical coordinates of the neighbourhoods using Python Geocoder package which return the latitude and longitude coordinates of each District. After that I am adding new data from Berlins official statistic Institute (importing them as excel) by merging different dataframes and bringing them togehther into one final dataset.

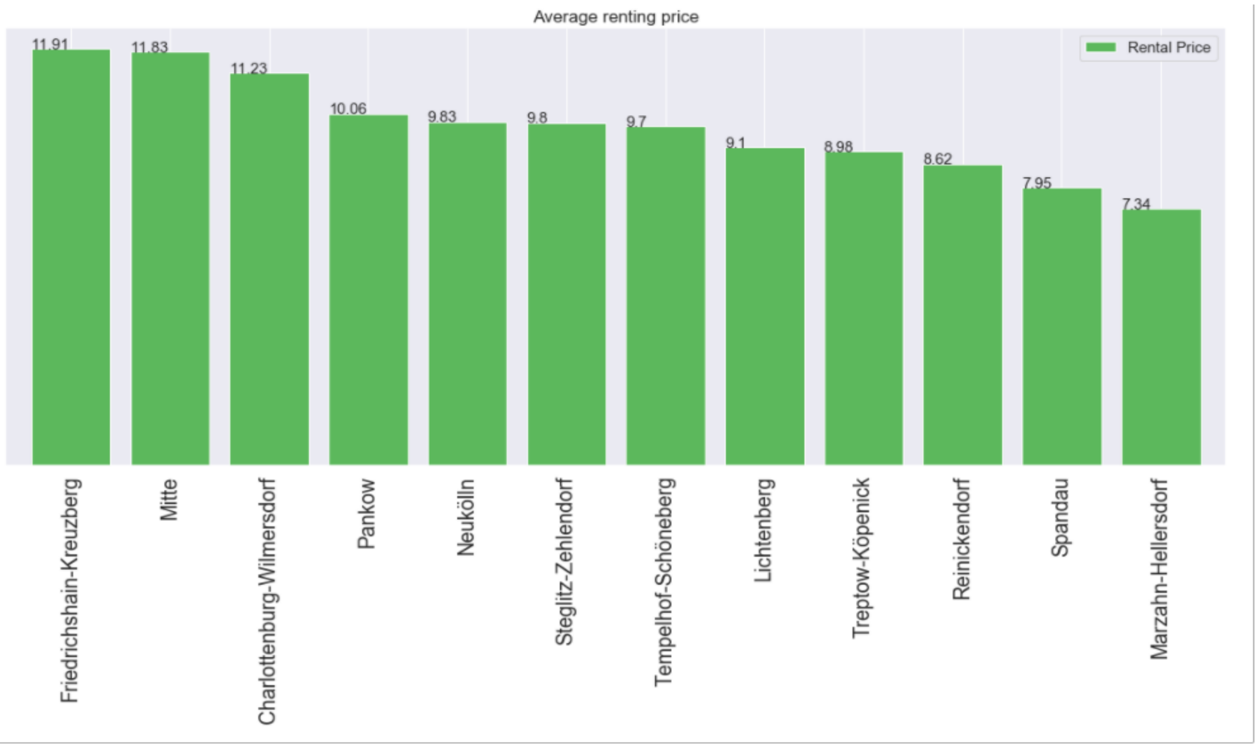
There were certain pre processing steps in order to bring the data into one final DataFrame and do the calculations etc.

The first part of the preprocessing process will be extracting the data regarding ‚resident location‘ and transform the data into percentages, as the given numbers represent total numbers which shall be standardized to the total population of the district. Moreover, I added information on the rental housing prices for each district (from CSV) as they aim to be the target variable in the regression model.

First assumptions show that residental location can only party explain differences in renting prices. Ditricts with high amount of good residental areas like Charlottenburg-Wilmersdorf or Steglitz-

Zehlendorf and Tempelhof- Schöneberg. Those Neighbourhoods have the biggest proportion of a good residental location which would explain high rental prices in that districts. But looking at the average rental prices per squaremeter in each district, it appears that districts which a high percentage of poor residental locations like Friedrichshain-Kreuzberg and Neukölln show the hightest renting prices. How can that be?



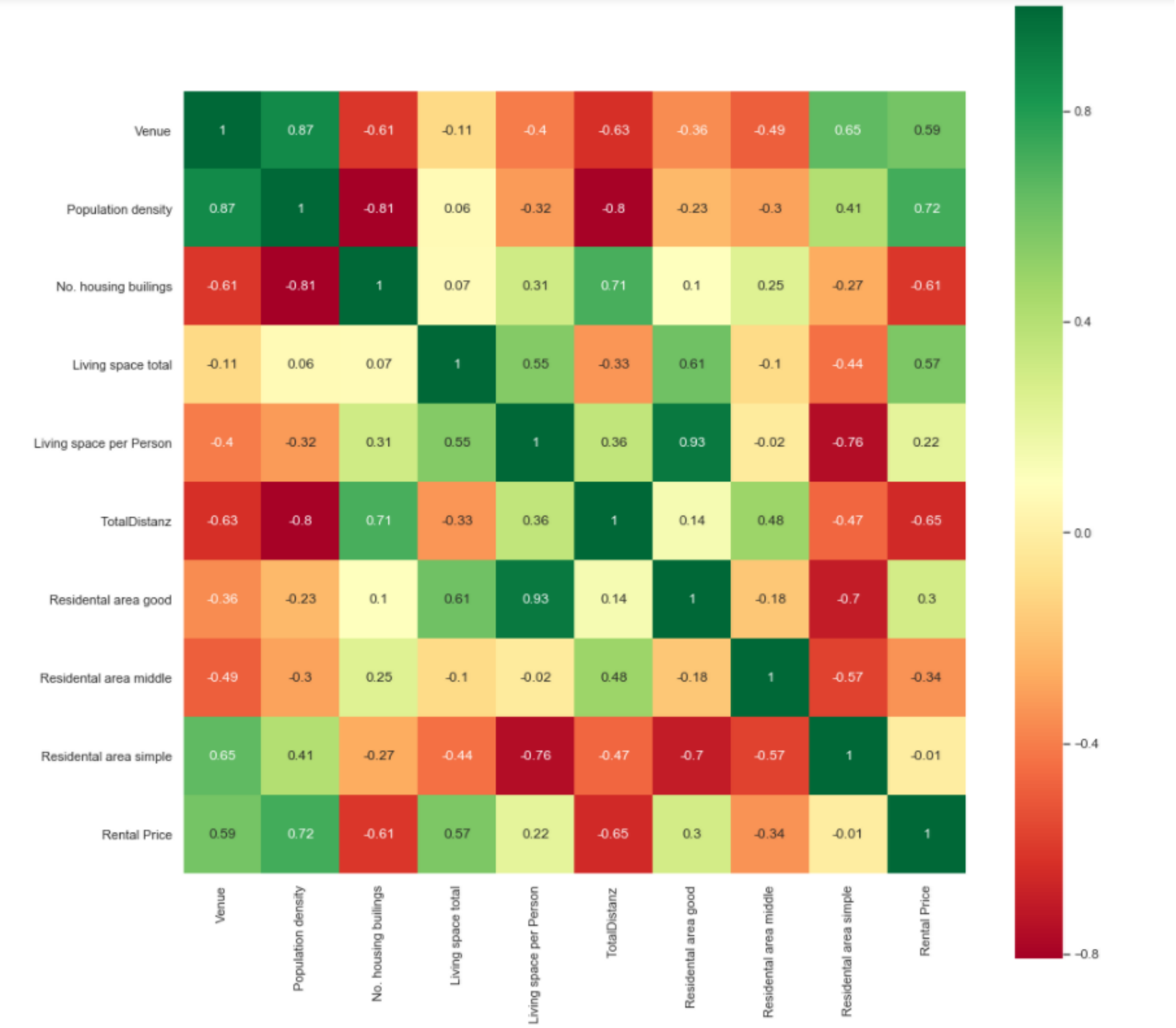


5. Results

Brining all the data together into a pandas Dataframe allowes me to do some statical calculations.

5.1. Correlation and Linear Regression

For the linear regression the first step was to check which variables have a high correlation with ‘rental price’ as the variable aims to be the target variable in the model.
Therefore a correlation- heatmap was created.



Looking at the correlation values and putting them in a scatterplot shows that only variable who tend to have a strong correlation should be used in the regression model:

- Venues
- Population density
- No. Of housing Buildings
- Distanz

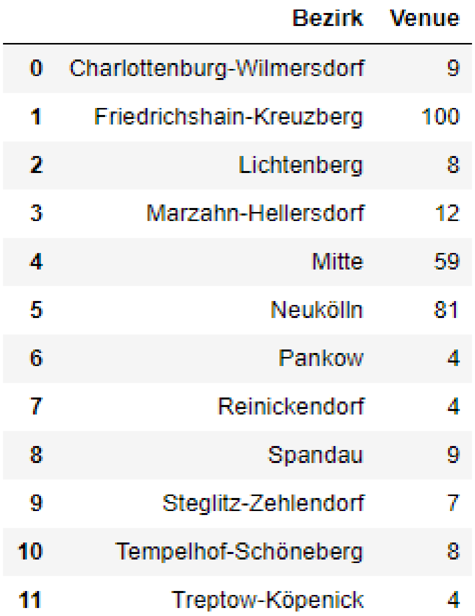
Supringsly the variables for the determination of the quality of the residental location did not show strong correlations with rental price. Therefore there will be left out in the regression model.

> Rental price = 4.1 + 2.19083196e-02 x ‚Venues‘ + 2.96047221e-05 x ‚Popolation density‘ - 8.41350337e-05 x No. Housing builidungs + 5.15239543e-04 x Living space + 7.53504401e+00 x Total Distance

R^2 is value that can tell us how good the modell is, it defines the amount of variance in the target: so about 90% can be explained by our variables

5.2 Cluster analysis

Another way to approach the data and to examine how rental prices are distributed in Berlin city was to look at the categories and amount of certain venues in each district. Using the Foursquare data, it was possible to determine how many and which Top 10 venues are to be found in each district. Accordingly, it was possible to examine the extent to which the number of venues in a district influences the average rental price of the district.



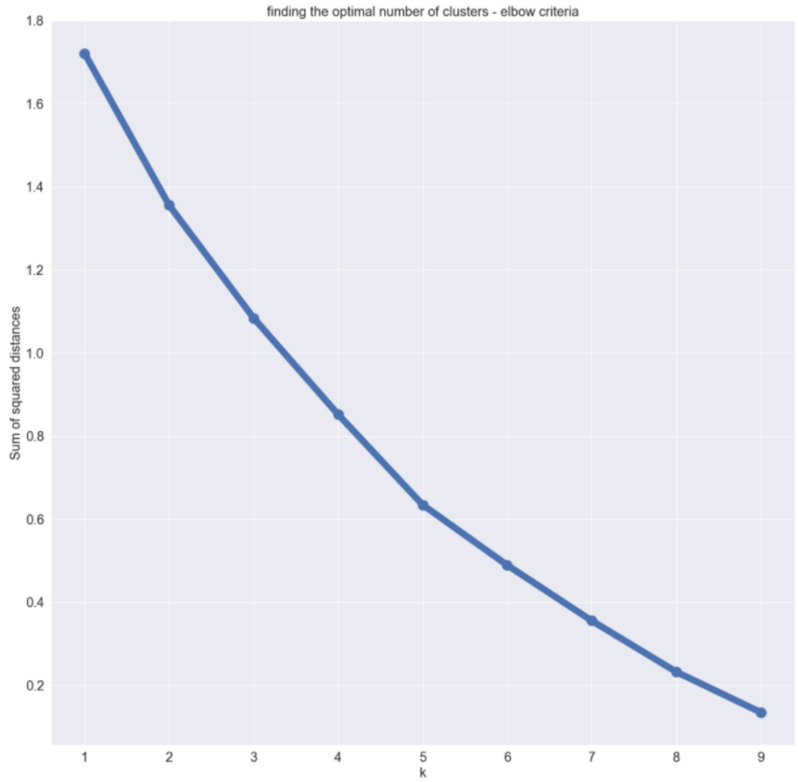
	Bezirk	Venue
0	Charlottenburg-Wilmersdorf	9
1	Friedrichshain-Kreuzberg	100
2	Lichtenberg	8
3	Marzahn-Hellersdorf	12
4	Mitte	59
5	Neukölln	81
6	Pankow	4
7	Reinickendorf	4
8	Spandau	9
9	Steglitz-Zehlendorf	7
10	Tempelhof-Schöneberg	8
11	Treptow-Köpenick	4

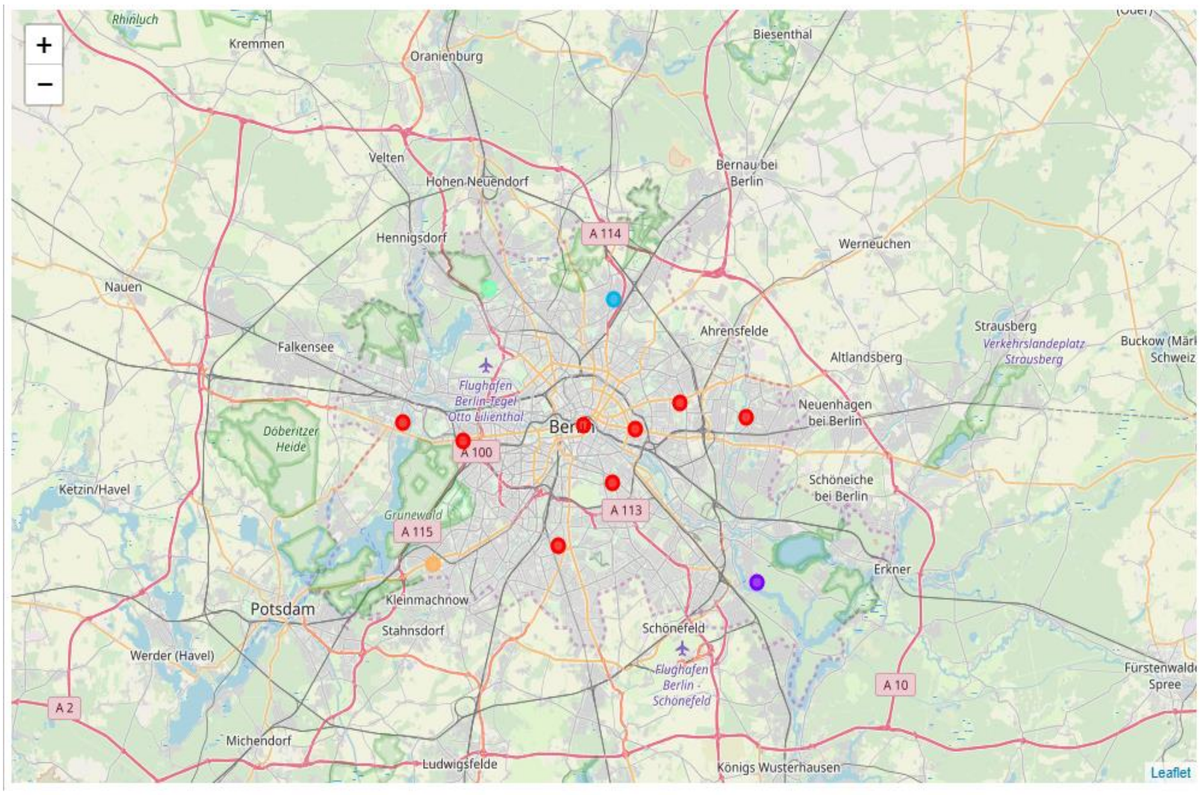
The table shows that the number of venues per district is reflected in the rents and thus the assumption can be made that there is a connection in this respect.

Via a cluster analysis certrain neighbourhoods were identified which are similiar along the Foursquare data. Herefore the correct number of cluster was determined by the so calles ‘elbow-criteria’. As a cluster means that datapoint within that cluster are similiar to each other the distance is used as a perimeter to calculate the smilarity. To avoid bigger influensce of larger numbers we standardized all variables with standard scalar.

Then I used the sum of the squared distance to determine the correct number of cluster (k=5).

As the elbow is seen at k=5, I will go for 5 Cluster. The results will be shown in the map of Berlin.





6. Conclusion

The previous analysis has shown a way on how to approach a dataset. First with a machine learning approach (multiple linear regression) then a cluster analysis. The linear regression gives us an easy way to create a model to describe – based on core features. It shows important factors that can influence renting prices in Berlin (or rental prices in general in big cities like Berlin). But of course important factors and variable were also left out and not included in the model. Moreover, when looking at the data it needs to be kept in mind, that tha explanatory variables wihtin the regression model might show high mulit-correlarity. For further analysis, other characteristics can and must of course be taken into consideration. Statistical statements in connection with the model could be better, but a detailed analysis in the context of this work was not feasible (due to lack of data). Moreover, the question of causality in this model was left open. Are the rents rising because there are many venues there or are there many venues because the high rents suggest that it is a popular area.

The cluster analysis teaches us how similiar dsitricts can be grouped based on different features. Nevertheless shows the clusteranalysis the typical divide of a city and confirms therefore what was to be expected. Central areas are more similar than districts that are more far ways. Cluster analysis also leaves room for further analysis. According to this, it would also be possible to look at the neighbourhoods on the basis of their inhabitant structure and examine who lives where and extract certain milieus.